DOCUMENT RESUME

ED 156 715 TM 007 249

AUTHOR	Levy, Paul S.; French, Dwight K.
TITLE	Synthetic Estimation of State Health Characteristics Based on the Health Interview Survey. Vital and Health Statistics. Series 2. Data Evaluation and Methods Research. Number 75.
INSTITUTION	National Center for Health Services Research (DHEW/PHS), Hyattsville, Md.
REPORT NO	DHEW-PHS-78-1349
PUB DATE	Oct 77
NOTE	30p.
AVAILABLE FROM	Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402 (Stock Number 260-937:35, $1.20)
EDRS PRICE	MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS	Census Figures; Community Health; Data Collection; Demography; Error Patterns; *Geographic Regions; Mathematical Models; National Surveys; Physical Health; *Public Health; *Reliability; Research Design; *Sampling; State Surveys; *Statistical Analysis; Statistical Bias; *Statistical Surveys
IDENTIFIERS	*Synthetic Estimation

ABSTRACT

Synthetic estimation is a statistical technique that estimates small-area statistics by combining national estimates of the relevant characteristics with estimates of other known characteristics of the small geographic area. The advantages of the synthetic estimation approach to local estimation are its intuitive appeal, its simplicity, and its low cost. A major disadvantage is its possibile lack of sensitivity to certain local characteristics. Another method used for the same purpose is the "nearly unbiased" estimator. The mathematics of both methods are presented, including formulas for estimating mean square error and average mean square error. An evaluation of synthetic estimates is demonstrated by comparing the results of a 50-cell grid with the results from 2, 4, 8, and 16-cell grids. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED156715

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Library of Congress Cataloging in Publication Data

United States. National Center for Health Statistics.
Synthetic estimation of State health characteristics based on the health interview survey.

(Vital and health statistics: Series 2, Data evaluation and methods research; no. 75) (DHEW publication; (PHS) 78-1349)
Bibliography
1. Health surveys—Statistical methods. 2. Estimation theory. 3. Health surveys—United States—States. I. Title. II. Series: United States. National Center for Health Statistics. Vital and health statistics: Series 2, Data evaluation and methods research; no. 75. III. Series: United States. Dept. of Health, Education, and Welfare. DHEW publication; (PHS) 78-1349. RA409.U45 no. 75 312'.07'23s [614.4'2] 77-22219
ISBN 0-8406-0107-7

For sale by the Superintendent of Documents, U.S. Government Printing Office

DATA EVALUATION AND METHODS RESEARCH

Series 2
Number 75

Synthetic Estimation of State Health Characteristics Based on the Health Interview Survey

This report discusses the various methods that have been proposed or used for obtaining estimates of health characteristics for local areas. Particular emphasis is given to discussion and evaluation of synthetic estimation procedures developed originally at the National Center for Health Statistics for purposes of estimating levels of health characteristics obtained from the Health Interview Survey for each State and the District of Columbia.

DHEW Publication No. (PHS) 78-1349

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Public Health Service
National Center for Health Statistics
Hyattsville, Md. October 1977

NATIONAL CENTER FOR HEALTH STATISTICS

DOROTHY P. RICE, *Director*

ROBERT A. ISRAEL, *Deputy Director*
JACOB J. FELDMAN, Ph.D., *Associate Director for Analysis*
GAIL F. FISHER, *Associate Director for the Cooperative Health Statistics System*
ELIJAH L. WHITE, *Associate Director for Data Systems*
JAMES T. BAIRD, JR., Ph.D., *Associate Director for International Statistics*
ROBERT C. HUBER, *Associate Director for Management*
MONROE G. SIRKEN, Ph.D., *Associate Director for Mathematical Statistics*
PETER L. HURLEY, *Associate Director for Operations*
JAMES M. ROBEY, Ph.D., *Associate Director for Program Development*
PAUL E. LEAVERTON, Ph.D., *Associate Director for Research*
ALICE HAYWOOD, *Information Officer*

OFFICE OF THE ASSOCIATE DIRECTOR FOR MATHEMATICAL STATISTICS

MONROE G. SIRKEN, Ph.D., *Associate Director*

Vital and Health Statistics-Series 2-No. 75

DHEW Publication No. (PHS) 78-1349
Library of Congress Catalog Card Number 77-22219

PREFACE

The investigations on which this report is based have been supported by contracts between the Measurement Research Laboratory, Office of Research, National Center for Health Statistics (NCHS), and the School of Public Health, University of Illinois at the Medical Center, Chicago. Dr. Monroe G. Sirken, Associate Director for Mathematical Statistics, NCHS, served as project officer for the Center on these investigations. Dr. Sirken and Ms. Patricia Royston, mathematical statistician, Office of Mathematical Statistics, NCHS, provided considerable input throughout the course of this project. Dr. Wesley L. Schaible and Dr. Dwight Brock, Office of Research, NCHS, reviewed an earlier draft of this report and made many helpful suggestions.

CONTENTS

Preface iii

Introduction 1

Review of the Literature 2

Alternative Methods of Obtaining State Estimates 3
- Background 3
- Nearly Unbiased Estimator 3
- Synthetic Estimator 6
- Regression-Adjusted Estimator 11

Evaluation of Synthetic Estimates 12
- Background 12
- Estimation of the Mean Square Error (MSE) and Average Mean Square Error (AMSE) of Synthetic Estimates 12
- Evaluation of HIS Synthetic Estimates for Alternative α-Cell Grids 13

References 16

Appendix: Proofs of Lemmas and Theorems 17

LIST OF TEXT TABLES

A. Interpretation of the components of the square of the bias of the "nearly" unbiased estimate 5

B. Distribution of percentage absolute differences between the nearly unbiased estimate and the true value among 42 States in the North, Central, South, and West Regions for total deaths, deaths from major cardiovascular-renal diseases, and deaths from motor vehicle accidents, 1960 6

C. Maximum contribution to the relative variance of $\tilde{\bar{X}}_s$ of sampling variation in the $P_{s\alpha}$ 10

D. Mean proportional absolute differences between the synthetic estimate produced by the total 50 α-cell grid and that produced by other α-cell grids for selected Health Interview Survey (HIS) variables, 1969-71 14

E. Maximum proportional absolute differences between the synthetic estimate produced by the total 50 α-cell grid and that produced by other α-cell grids for selected Health Interview Survey (HIS) variables, 1969-71 14

F. Correlation coefficients between the synthetic estimate produced by the total 50 α-cell grid and that produced by other α-cell grids for selected Health Interview Survey (HIS) variables, 1969-71 15

SYNTHETIC ESTIMATION OF STATE HEALTH CHARACTERISTICS BASED ON THE HEALTH INTERVIEW SURVEY

Paul S. Levy, Sc.D., *School of Public Health, University of Illinois at the Medical Center, Chicago,* and Dwight K. French, *Statistical Methods Staff, National Center for Health Statistics*

INTRODUCTION

Statisticians, demographers, economists, and others have long been aware of the critical need for accurate small area statistics. While the U.S. decennial census provides accurate local statistics of many characteristics once every 10 years, the accuracy of these statistics becomes questionable as time elapses from the last census and, in addition, characteristics other than those found on the census questionnaire are often desired.

Although a rather extensive system of ongoing general purpose surveys is conducted by Federal agencies, they are almost always designed to produce estimates for the United States as a whole or, at most, for rather large geographic regions or divisions. For reasons of sample size and design, direct estimates for such subdivisions as cities, counties, States, or other minor civil divisions, which are so critically needed, can rarely be obtained from these surveys.

The National Center for Health Statistics, one of the Federal agencies responsible for maintaining a system of sample surveys and other data collection systems, has long recognized the need for good small area statistics, and for the past decade has investigated alternate strategies for obtaining such estimates. In particular, NCHS has developed a procedure known as "synthetic estimation" for obtaining small area statistics. This procedure obtains small area estimates of characteristics by combining national estimates of the characteristics specific to demographic subgroups with estimates of the proportional distribution of the local population into the subgroups. The subgroups would be chosen for their relevance to the characteristic being estimated. For example, if it were desired to estimate the prevalence of the sickle cell trait in a particular county having a racial distribution of 30 percent white and 70 percent black, and if a hypothetical national survey estimated that the trait was prevalent among 10 percent of U.S. blacks and virtually nonexistent among U.S. whites, one would estimate that 7 percent of the population in the county had the trait (30% × 0% + 70% × 10%). This is a synthetic estimate.

The advantages of the synthetic-estimation approach to local estimation are its intuitive appeal, its simplicity, and its low cost relative to a direct survey of the local population. A major disadvantage is its lack of sensitivity to certain local characteristics. For example, in the above illustration, it may happen that the white population in that area are all of Mediterranean descent and have more than a negligible amount of persons with the sickle cell trait.

Much research on the synthetic estimation procedure has emerged since an NCHS report on synthetic estimates of disability for States was published in 1968.[1] The purpose of this report is to examine critically the various methods for

obtaining local estimates that are in the literature and, in particular, to examine synthetic estimation from a methodological point of view.

REVIEW OF THE LITERATURE

The need for methods of obtaining valid and reliable estimates of characteristics of local populations has been recognized for a long time by statisticians and demographers. In particular, much effort has been expended by statisticians associated with the U.S. Bureau of the Census and their contractors, especially in the use of *symptomatic variables* such as births, deaths, and school enrollment, which are available on a local level, to measure changes in population size since the most recent decennial census. Methods such as the *vital rates technique, censal ratio method, Census Bureau Component Methods I and II, ratio correlation method,* and others have been described extensively in the literature.[2] Basically, these methods use the relationship between the population size of the local area at the most recent census and the measure of the symptomatic variable or variables for that year, in conjunction with the value of the symptomatic variable(s) at the date for which the estimate is desired, to produce the desired local estimate of population size or change. An elaboration of the use of techniques based on symptomatic variables has been developed recently by Ericksen.[3-5] His elaboration involves use of sample data from the Current Population Survey in conjunction with symptomatic variables to obtain estimates of population size for local areas.

Although health statisticians have long felt the need for valid and reliable estimates of health characteristics for local areas, only in the past decade has serious attention been given to the development of methodology for obtaining local area estimates of such health characteristics as morbidity, mortality, disability, and utilization of health care services. The methods developed by demographers for estimating local population sizes could not, however, be directly applied to the estimation of health characteristics for local areas; hence, methodology for estimating health conditions for local areas has developed along different lines from those discussed above for local estimation of population size.

A major advance in estimation of health characteristics for local areas came with an experiment conducted by Walt R. Simmons and his staff at the National Center for Health Statistics (NCHS) during the mid-1960's and published in 1968.[1] In this experiment, three different estimation techniques were used to produce estimates of long- and short-term disability for each State in the United States for the 2-year period beginning July 1, 1962, and ending June 30, 1964. The NCHS data used for estimating disability were from the Health Interview Survey (HIS), and the population data were from the Current Population Survey update of the 1960 Decennial Census.

One of the methods used was proposed originally by Woodruff to produce local estimates of retail trade;[6] the other two, namely, the synthetic estimator and the nearly unbiased estimator, were developed at NCHS. These methods will be discussed in greater detail. Of the three methods investigated, the synthetic estimator was judged to be the most promising for estimating disability on the State level, and the estimates finally published[1] were obtained by this method.

The NCHS publication on synthetic estimates of disability[1] seemed to stimulate further efforts to apply and evaluate synthetic estimation. Within NCHS, an evaluation of the synthetic estimation procedure was conducted in which synthetic estimates of death rates in 1960 from four causes (motor vehicle accidents, major cardiovascular-renal diseases, suicide, and tuberculosis) were calculated for each State and for the District of Columbia.[7] These synthetic estimates of death rates were then compared to the known true death rates for each State and, in general, agreement between synthetic estimates and true death rates was good for one of the causes examined (major cardiovascular-renal diseases), fair for another (suicides), and poor for the other two (motor vehicle accidents and tuberculosis). The general conclusion from the study was that the validity and reliability of synthetic estimates might differ from characteristic to characteristic.

As part of the NCHS study of death rates, an

alternative estimation procedure was developed. The resulting estimator, called the regression-adjusted estimator, uses the synthetic estimate in combination with ancillary data available on the State level and thought to be correlated with the health characteristic to be estimated.[7] This estimate, for at least one of the causes of death examined, seemed to be an improvement over the synthetic estimator.

After the NCHS publication on synthetic estimates of disability, the Bureau of the Census produced synthetic estimates of unemployment rates and number of dilapidated housing units that had all plumbing facilities for States, SMSA's, and counties.[8-10] In addition, extensive studies were undertaken to evaluate the synthetic estimates. An important result of these studies was the emergence of a criterion, called the average mean square error (AMSE), as a proposed measure of the accuracy of a set of synthetic estimates, and the development of a method for estimating the AMSE.[9,11] These methods will be discussed in greater detail later in this report.

Most recently, Namekata, Levy, and O'Rourke[12] investigated the use of synthetic estimation in obtaining estimates of complete and partial work disability for States based on data from the 1970 census. The synthetic estimates were obtained and compared with the direct estimates that were available from the 1970 Decennial Census for each State. Their general conclusions were that the synthetic estimation technique was fairly good for partial work disability but fairly poor for complete work disability.

ALTERNATIVE METHODS OF OBTAINING ESTIMATES

Background

In the original NCHS investigation of alternative procedures for small area estimation of health characteristics from the Health Interview Survey (HIS), several procedures were considered.[1] In this section, we will discuss in detail two of the methods, namely, the nearly unbiased estimator and the synthetic estimator. In addition, we will discuss an estimator, called the regression-adjusted estimator, not considered in the original investigation but developed in a later study.[7]

One of the problems in obtaining estimates for States of health characteristics based on HIS data is the fact that the basic design of the HIS does not lend itself to unbiased estimates for States. In the basic HIS design, a primary sampling unit (PSU), which is generally a county or SMSA, is chosen to represent a stratum consisting of one or more demographically similar PSU's. Those strata consisting of more than one PSU are called non-self-representing strata, and their component PSU's may not be from the same State, although they would be from the same census region (Northeast, North Central, South, or West). Thus, the estimate from a sample PSU when inflated to represent the entire stratum might cut across State boundaries, and hence, it would be impossible to combine the unbiased estimates for strata into unbiased estimates for States.

Nearly Unbiased Estimator

One of the methods considered in the original NCHS investigation is called the nearly unbiased estimator and yields an estimate for a State that is technically nearly unbiased. Basically, this procedure takes the usual HIS stratum estimate for an aggregate and allocates it to a State in relation to the proportion of the total stratum population coming from the State. In other words, the nearly unbiased estimate $\bar{X}'_s$ of the mean level of characteristic X for State s is given by

$$\bar{X}'_s = \sum_{j=1}^{J} \frac{n_{sj.}}{n_{s..}} \bar{X}'_j \qquad (1)$$

where

$\bar{X}'_j$ = unbiased HIS estimate for the mean level of X for stratum j,

$n_{s..} = \sum_{j=1}^{J} n_{sj.}$

= the number of persons in State s,

$$n_{sj.} = \sum_{i=1}^{I_{sj}} n_{sji}$$

= the number of persons in that portion of stratum j which is in State s,

n_{sji} = the number of persons in the jth stratum, sth State, ith PSU; $s=1,\ldots, S$; $j=1,\ldots, J$; $i=1,\ldots, I_{sj}$, and

I_{sj} = the number of PSU's in stratum j that are in State s.

Some properties of the nearly unbiased estimate $\bar{X}'_s$ are given below. The proofs are presented in appendix I.

Lemma 1: The expectation $E(\bar{X}'_s)$ of the nearly unbiased estimator $\bar{X}'_s$ is given by

$$E(\bar{X}'_s) = \sum_{j=1}^{J} \frac{n_{sj.}\bar{X}_{.j.}}{n_{s..}} \qquad (2)$$

where

$$\bar{X}_{.j.} = \sum_{s=1}^{S} \frac{n_{sj.}\bar{X}_{sj.}}{n_{.j.}}$$

= average level of X in stratum j,

$$n_{.j.} = \sum_{s=1}^{S} n_{sj.}$$

= number of persons in stratum j,

$$\bar{X}_{sj.} = \sum_{i=1}^{I_{sj}} \frac{n_{sji}\bar{X}_{sji}}{n_{sj.}}$$

= average level of X in that portion of stratum j which is in State s, and

$\bar{X}_{sji}$ = average level of characteristic X in that portion of State s which is in PSU i of stratum j.

Lemma 2: The bias $B(\bar{X}'_s)$ in the nearly unbiased estimate $\bar{X}'_s$ can be expressed by

$$B(\bar{X}'_s) = \sum_{j=1}^{J} \frac{n_{sj.}(\bar{X}_{.j.} - \bar{X}_{sj.})}{n_{s..}} \qquad (3)$$

or equivalently by

$$B(\bar{X}'_s) = \sum_{j=1}^{J} \frac{n_{sj.}}{n_{s..}} \sum_{i=1}^{I_{sj}} \left(\frac{\bar{X}_{.j.}}{I_{sj}} - \frac{n_{sji}}{n_{sj.}} \bar{X}_{sji} \right). \qquad (4)$$

Lemma 3: Let us assume that the ratio $n_{sji}/n_{sj.}$ is the same for all PSU's in the same State, which implies that

$$\frac{n_{sji}}{n_{sj.}} = \frac{1}{I_{sj}} \text{ for } i=1, \ldots, I_{sj}. \qquad (5)$$

Then the bias $B(\bar{X}'_s)$ in the nearly unbiased estimate $\bar{X}'_s$ has the form given by

$$B(\bar{X}'_s) = \sum_{j=1}^{J} \frac{n_{sj.}}{n_{s..}} \frac{1}{I_{sj}} \sum_{i=1}^{I_{sj}} (\bar{X}_{.j.} - \bar{X}_{sji}) \qquad (6)$$

Theorem 4: If $\frac{n_{sji}}{n_{sj.}} = \frac{1}{I_{sj}}$ for $i=1,\ldots, I_{sj}$, then $B^2(\bar{X}'_s)$, the square of the bias in $\bar{X}'_s$, is given by the expression

$$B^2(\bar{X}'_s) = \sum_{j=1}^{J} \frac{n_{sj.}^2}{n_{s..}^2} \frac{\sigma_{sji}^2}{I_{sj}} + \sum_{j=1}^{J} \frac{n_{sj.}^2}{n_{s..}^2} \frac{1}{I_{sj}} (\bar{X}_{sj.} - \bar{X}_{.j.})^2 + 2 \sum_{j=1}^{J} \frac{n_{sj.}^2}{n_{s..}^2} \frac{1}{I_{sj}^2} \sum_{i=1}^{I_{sj}} \sum_{i'<i} (\bar{X}_{.j.} - \bar{X}_{sji})(\bar{X}_{.j.} - \bar{X}_{sji'})$$

$$+ 2 \sum_{j=1}^{J} \sum_{k<j} \frac{n_{sj.} n_{sk.}}{n_{s..}^2 I_{sj} I_{sk}} \sum_{i=1}^{I_{sj}} (\bar{X}_{.j.} - \bar{X}_{sji}) \sum_{i=1}^{I_{sk}} (\bar{X}_{.k.} - \bar{X}_{ski}), \qquad (7)$$

Table A. Interpretation of the components of the square of the bias of the "nearly" unbiased estimate

Component	Interpretation
1. $\sum_{j=1}^{J} \frac{n_{sj.}^2}{n_{s..}^2 I_{sj}} (\bar{X}_{sj.} - \bar{X}_{j.})^2$	Represents difference between the average level of X for a stratum and the average level of X for the portion of the straum that is in State s.
2. $\sum_{j=1}^{J} \frac{n_{sj.}^2}{n_{s..}^2} \frac{\sigma_{sji}^2}{I_{sj}}$	Represents variance in X among PSU's belonging to the same State and stratum.
3. $2\sum_{j=1}^{J} \sum_{k<j} \frac{n_{sj.} n_{sk.}}{n_{s..}^2 I_{sj} I_{sk}} \sum_{i=1}^{I_{sj}} (\bar{X}_{j.} - \bar{X}_{sji}) \sum_{i=1}^{I_{sk}} (\bar{X}_{k.} - \bar{X}_{ski})$	Represents a "between-strata" covariance.
4. $2\sum_{j=1}^{J} \frac{n_{sj.}^2}{n_{s..}^2} \frac{1}{I_{sj}^2} \sum_{i=1}^{I_{sj}} \sum_{i'<i} (\bar{X}_{sji} - \bar{X}_{j.})(\bar{X}_{sji'} - \bar{X}_{j.})$	Represents a "between-PSU" covariance.

where

$$\sigma_{sji}^2 = \frac{1}{I_{sj}} \sum_{i=1}^{I_{sj}} (\bar{X}_{sji} - \bar{X}_{sj.})^2 .$$

Theorem 4 implies that under the condition that $n_{sji}/n_{sj.}$ is the same for all PSU's within the same stratum and belonging to the same State, the square of the bias of the nearly unbiased estimate $\bar{X}_s'$ consists of four components specified in table A.

It can be shown that the third component of $B^2(\bar{X}_s')$ can be transformed to the equivalent algebraic form given by

$$2\sum_{j=1}^{J} \sum_{k<j} \frac{n_{sj.} n_{sk.}}{n_{s..}^2} (\bar{X}_{sj.} - \bar{X}_{j.})(\bar{X}_{sk.} - \bar{X}_{k.}).$$

Thus, an equivalent expression for the square of the bias in the nearly unbiased estimate $\bar{X}_s'$ is given by

$$B^2(\bar{X}_s') = \sum_{j=1}^{J} \frac{n_{sj.}^2}{n_{s..}^2} \frac{\sigma_{sji}^2}{I_{sj}} + \sum_{j=1}^{J} \frac{n_{sj.}^2}{n_{s..}^2 I_{sj}} (\bar{X}_{sj.} - \bar{X}_{j.})^2 + 2\sum_{j=1}^{J} \sum_{k<j} \frac{n_{sj.} n_{sk.}}{n_{s..}^2} (\bar{X}_{sj.} - \bar{X}_{j.})(\bar{X}_{sk.} - \bar{X}_{k.})$$

$$+ 2\sum_{j=1}^{J} \frac{n_{sj.}^2}{n_{s..}^2 I_{sj}^2} \sum_{i=1}^{I_{sj}} \sum_{i'<i} (\bar{X}_{sji} - \bar{X}_{j.})(\bar{X}_{sji'} - \bar{X}_{j.}) \qquad (8)$$

Variance and mean square error of the nearly unbiased estimate.—The variance $\sigma_{\bar{X}_s'}^2$ can be obtained directly from its definitional formula. This is given by

$$\sigma_{\bar{X}_s'}^2 = \sum_{j=1}^{J} \frac{n_{sj.}^2}{n_{s..}^2} \sigma_{\bar{X}_j'}^2 \qquad (9)$$

where $\sigma_{\bar{X}_j'}^2$ is the variance of $\bar{X}_j'$, the unbiased estimate of the mean level of X in stratum j.

It follows that the mean square error $\mathrm{MSE}_{\bar{X}_s'}$ of the nearly unbiased estimate $\bar{X}_s'$ is given by

$$\mathrm{MSE}_{\bar{X}_s'} = \sigma_{\bar{X}_s'}^2 + B^2(\bar{X}_s')$$

where $\sigma_{\bar{X}_s'}^2$ is given by relation (9), and $B^2(\bar{X}_s')$ is given by relation (8) (under the condition that $n_{sji}/n_{sj.} = 1/I_{sj}$).

Evaluation of the nearly unbiased estimator.—In the original NCHS investigation of methods for obtaining local estimates, the nearly

unbiased estimate did not emerge as the method of choice for producing these local area HIS estimates of health characteristics primarily because examination of the estimates produced by this method showed evidence that they were unstable.[1]

A later study was performed at NCHS to determine the extent to which the nearly unbiased estimator might be biased. The data base chosen for this study was mortality data in 1960 for the 42 States in the North Central, South, and West Regions of the United States. In particular, nearly unbiased estimates of total deaths, deaths from motor vehicle accidents, and deaths from major cardiovascular-renal diseases were obtained from each State using the same stratification that is used in the Health Interview Survey. These nearly unbiased estimates were then compared with the true number of deaths in each State in the three regions examined, and the biases of the estimators were evaluated by the percentage absolute difference $\frac{100|\bar{X}'_s - \bar{X}_{s..}|}{\bar{X}_{s..}}$.

Table B. Distribution of percentage absolute differences between the nearly unbiased estimate and the true value among 42 States in the North Central, South, and West Regions for total deaths, deaths from major cardiovascular-renal diseases, and deaths from motor vehicle accidents, 1960

Percentage absolute difference between nearly unbiased estimate and true value	Cause of death		
	All causes	Major cardiovascular-renal diseases	Motor vehicle accidents
	Frequency		
Total	42	42	42
0.0-0.9	16	15	12
1.0-1.9	8	8	6
2.0-2.9	6	5	4
3.0-3.9	5	3	1
4.0-4.9	1	2	3
5.0-5.9	1	2	3
6.0-6.9	3	2	1
7.0-7.9	0	1	2
8.0-8.9	2	0	5
9.0-9.9	2	4	5
	Percent		
Median percentage absolute difference	1.78	1.70	2.70

The distribution of percentage absolute difference is given in table B. The median percentage absolute difference was 1.78 percent for total deaths, 1.70 percent for major cardiovascular-renal deaths and 2.70 percent for motor vehicle accident deaths. The small biases obtained from this empirical study would yield the interpretation that for the stratification used in HIS, the nearly unbiased estimator is in fact an estimator having small bias.

However, in a given year, the number of households interviewed in a particular stratum might be quite small for the Health Interview Survey. It is, therefore, anticipated that $\sigma^2_{\bar{X}_j}$, and hence $\sigma^2_{\bar{X}'_s}$, the sampling variance of the nearly unbiased estimator might be quite large. In addition, the $\sigma^2_{\bar{X}_j}$ are difficult to estimate from the data. Hence, in terms of sampling variance, the nearly unbiased estimate $\bar{X}'_s$ might not be the method of choice.

Synthetic Estimator

Background.—The other method for obtaining local estimates of health characteristics investigated in the original NCHS study[1] is known as synthetic estimation and was the method finally chosen for producing local estimates of HIS health characteristics for States in 1963-64 and again in 1969-71.[15] The underlying rationale for synthetic estimation is that the distribution of a health characteristic does not vary among populations of States except to the extent that States vary in demographic composition. In other words, the method assumes that the incidence or prevalence of a health characteristic would be the same for two States if their composition were the same with respect to such demographic variables as age, sex, race, family income, family size, place of residence, and industry of the head of the family.

Conceptually, synthetic estimation uses the model given by

$$\tilde{\bar{X}}_s = \sum_{\alpha=1}^{k} P_{s\alpha} \bar{X}_\alpha \qquad (10)$$

where

$\tilde{\bar{X}}_s$ = mean level of characteristic X for the sth State,

$P_{s\alpha}$ = proportion of the population who are members of population cell α (alpha), which is the socioeconomic demographically bounded class of specified age, sex, race, income, etc. The sum over all α cells of $P_{s\alpha}$ = unity.

$\bar{X}_\alpha$ = mean level of characteristic X for persons in cell α in the United States as a whole, and

k = number of α cells utilized.

In the original NCHS investigation, the $\bar{X}_\alpha$ were national estimates of HIS variables for the period July 1962-June 1964. The population estimates $P_{s\alpha}$ were obtained from tabulations of a 5-percent sample questionnaire of the 1960 Decennial Census of Population for the 50 States and the District of Columbia. The population α cells were defined by cross-classifications of the following variables:

1. Color: white; all other
2. Sex: male; female
3. Age group: under 17 years; 17-44 years; 45-64 years; 65 years and over
4. Residence: standard metropolitan statistical area (SMSA)-central city; SMSA-not central city; not SMSA
5. Family income: under $4,000; $4,000 and over
6. Family size: fewer than seven members; seven members or more
7. Industry of head of family: Standard Industrial Classification codes 1 through 17 (Forestry and Fisheries, Agriculture, and Construction) and codes 19 and over (All Other Industries)

The 384 possible cross-classification cells were collapsed to 78 so that reliable estimates could be obtained from the Health Interview Survey for each α cell.

For the synthetic estimates for the years 1969-71, HIS data from the three surveys of 1969, 1970, and 1971 were used to obtain the rates or percentages of the health characteristics measured. The populations of the 50 States and the District of Columbia were obtained from a sample described in a publication of the U.S. Bureau of the Census entitled *Public Use Samples of Basic Records From the 1970 Decennial Census, Description and Technical Documentation*, published in 1972. Of six such samples, the one used was the State Public Use Sample from the 5-percent questionnaires. Persons in the military or confined to institutions were not included in the population estimates produced for each State. Thus the restriction of the HIS samples to the civilian noninstitutionalized population was carried over to these synthetic estimates. Of the seven variables used to produce the 78 α cells for the original report,[1] only six were available in the Public Use Sample used to produce these synthetic estimates. The variable that was not available was residence in standard metropolitan statistical areas. The six variables can produce a possible 128 cells of data. These were collapsed to 50 α cells for which reliable national estimates from the Health Interview Survey could be provided. A regional adjustment (as specified below) was employed for the State estimates within each of the four geographic regions of the United States to make these estimates consistent with the regional estimates produced by the probability design of the Health Interview Survey.

In summary, the synthetic estimates produced for HIS health characteristics for the years 1969-71[13] use the same basic method as was used in the original NCHS investigation. However, in addition to estimates of long- and short-term disability, estimates of utilization of medical services were provided as well as estimates for subdomains of the population of each State (age, sex, color, and family income). Also, a methodology has been developed for providing sampling variances of synthetic estimates.

Detailed synthetic estimate.—The detailed synthetic estimate $\tilde{\bar{X}}_s$ of the mean level of characteristic X for State s is given by

$$\tilde{\bar{X}}_s = \bar{\bar{X}}_s \frac{\bar{X}'_r}{\sum_{t=1}^{T} \bar{X}_t P_{rt}} \qquad (11)$$

where

$\tilde{\bar{X}}_s$ = final synthetic estimate of the mean level of characteristic X for State s,

$\bar{X}'_r$ = the usual HIS final estimate of the mean level of characteristic X for region r (where region r contains State s),

P_{rt} = proportion of the population (from the 1970 Decennial Census) of region r which is in State t ($t=1,\ldots, T$),

T = number of States in region r,

$$\bar{\bar{X}}_s = \sum_{\alpha=1}^{k} P'_{s\alpha}\bar{X}'_\alpha$$

= first-stage synthetic estimate of characteristic X for State s,

$\bar{X}'_\alpha$ = final HIS estimate of the mean level of characteristic X for demographic cell α for the United States,

$P'_{s\alpha}$ = estimated proportion of the 1970 population in State s belonging to cell α (as estimated from the 1970 U.S. Census 1-Percent Public Use Tapes), and

k = number of α cells.

Synthetic estimates $\tilde{\bar{X}}_{su}$ for subdomains (age, sex, color, and income) are given by the estimator

$$\tilde{\bar{X}}_{su} = \bar{\bar{X}}_{su}\frac{\tilde{\bar{X}}_s}{\sum_u f_{su}\bar{\bar{X}}_{su}} \qquad (12)$$

where

$\tilde{\bar{X}}_{su}$ = the final synthetic estimate for the mean level of characteristic X for subdomain u within State s,

$$\bar{\bar{X}}_{su} = \sum_{\alpha \in u} P'_{s\alpha}\bar{X}'_\alpha / f_{su}$$

= preliminary synthetic estimate for the mean level of characteristic X in subdomain u of State s,

$P'_{s\alpha}$ = the estimated proportion (as estimated for the U.S. Census 1970 1-Percent Public Use Sample Tapes) of the population of State s belonging to cell α, and

$$f_{su} = \sum_{\alpha \in u} P'_{s\alpha}$$

= the estimated proportion of the population of State s belonging in subdomain u, except for synthetic estimates of work-loss days per person per year in age groups. By definition, HIS excludes all persons under 17 years of age from the employed population. Therefore, the factor f_{su} in the denominator of the ratio adjustment for these statistics is redefined as the estimated proportion of the population age 17 and over in State s that belongs to subdomain (age group) u.

The synthetic estimates for subdomains as given by equation (12) are ratio adjusted so that the aggregates are consistent with the final synthetic estimates for the State as a whole.

The α variables were limited to those listed below:

Color: white; other than white.

Sex: male; female.

Age group: under 17 years; 17-44 years; 45-64 years; 65 years and over.

Family income: under $5,000; $5,000 and over.

Family size: fewer than seven members; seven members or more.

Industry of head of family: Standard Industrial Classification Codes 1 through 17 (agriculture, forestry, fisheries, mining, and construction); 18 and above (all others).

The 128 cross-classification cells produced by these variables were collapsed into 50 α cells for which reliable national estimates from the Health Interview Survey could be made.

The ratio adjustment $\bar{X}'_r / \sum_{t=1}^{T} P_{rt}\bar{\bar{X}}_t$ was included in order to reflect a regional component in final estimates. It is the ratio of the published regional figure to the preliminary, derived regional rate calculated from the State estimates.

Estimation of sampling errors of synthetic estimates for HIS characteristics 1969-71.—The

synthetic estimates presented for the 1969-71 HIS data are subject to sampling variability from two sources because they are based on HIS estimates and estimated population proportions. (When synthetic estimates are computed from known proportions and population means they are not subject to sampling error, since there would be no sampling involved in the synthetic estimation procedure.) The sampling variance of a synthetic estimate $\tilde{\bar{X}}_s$ (ignoring the regional adjustment) is given by

$$\sigma^2_{\tilde{\bar{X}}_s} = \mathrm{Var}\left(\sum_{\alpha=1}^{k} P'_{s\alpha}\bar{X}'_\alpha\right) = \sum_{\alpha=1}^{k} \mathrm{Var}\,(P'_{s\alpha}\bar{X}'_\alpha) + 2\sum_{\alpha<\alpha'} \mathrm{Cov}\,(P'_{s\alpha}\bar{X}'_\alpha, P'_{s\alpha'}\bar{X}'_{\alpha'}) \quad (13)$$

Theorem 5: If the $P'_{s\alpha}$ are independent of the $\bar{X}'_\alpha$, then the variance of $\tilde{\bar{X}}_s$ given by equation (13) reduces to

$$\sigma^2_{\tilde{\bar{X}}_s} \doteq \sum_{\alpha=1}^{k} P^2_{s\alpha}\sigma^2_{\bar{X}'_\alpha} + \frac{1}{n_s}\sum_{\alpha=1}^{k} \bar{X}^2_\alpha (1 - P_{s\alpha})P_{s\alpha} + 2\sum_{\alpha<\alpha'} P_{s\alpha}P_{s\alpha'}\,\mathrm{Cov}\,(\bar{X}'_\alpha, \bar{X}'_{\alpha'}) \quad (14)$$

for large values of n_s.

The first and third terms of equation (14) represent the variance of $\tilde{\bar{X}}_s$ if the $P'_{s\alpha}$ were not subject to sampling variation. Thus, the effect of sampling variation of the $P'_{s\alpha}$ on $\sigma^2_{\tilde{\bar{X}}_s}$ is measured by the expression

$$\frac{1}{n_s}\sum_{\alpha=1}^{k} \bar{X}^2_\alpha P_{s\alpha}(1 - P_{s\alpha}). \quad (15)$$

If P^m_s = minimum $(P_{s1}, \ldots, P_{sk})$ and $V^2_{\tilde{\bar{X}}_s}$ = Rel-variance of $\tilde{\bar{X}}_s$, we have

$$V^2_{\tilde{\bar{X}}_s} = \frac{\sum_{\alpha=1}^{k} P^2_{s\alpha}\sigma^2_{\bar{X}'_\alpha} + 2\sum_{\alpha'<\alpha} P_{s\alpha}P_{s\alpha'}\,\mathrm{Cov}\,(\bar{X}'_\alpha, \bar{X}'_{\alpha'})}{[E(\bar{X}_s)]^2} + \frac{\frac{1}{n_s}\sum_{\alpha=1}^{k} \bar{X}^2_\alpha P_{s\alpha}(1 - P_{s\alpha})}{[E(\tilde{\bar{X}}_s)]^2}.$$

But

$$\sum_{\alpha=1}^{k} \bar{X}^2_\alpha P_{s\alpha}(1 - P_{s\alpha}) = \sum_{\alpha=1}^{k} \bar{X}^2_\alpha P^2_{s\alpha}\frac{1 - P_{s\alpha}}{P_{s\alpha}} \leq \frac{1 - P^m_s}{P^m_s}\sum_{\alpha=1}^{k} (\bar{X}_\alpha P_{s\alpha})^2 \leq \frac{1 - P^m_s}{P^m_s}\left[\sum_{\alpha=1}^{k} P_{s\alpha}\bar{X}_\alpha\right]^2,$$

and

$$E(\tilde{\bar{X}}_s) = \sum_{\alpha=1}^{k} P_{s\alpha}\bar{X}_\alpha\ ;$$

therefore,

$$\sum_{\alpha=1}^{k} \bar{X}^2_\alpha P_{s\alpha}(1 - P_{s\alpha}) \leq \frac{1 - P^m_s}{P^m_s}\,[E(\tilde{\bar{X}}_s)]^2$$

and the rel-variance $V^2_{\tilde{\bar{X}}_s}$ of the synthetic estimate $\tilde{\bar{X}}_s$ satisfies the inequality given by

$$V^2_{\tilde{\bar{X}}_s} < \frac{\sum_{\alpha=1}^{k} P^2_{s\alpha}\sigma^2_{\bar{X}'_\alpha} + 2\sum_{\alpha'<\alpha} P_{s\alpha}P_{s\alpha'}\,\mathrm{Cov}\,(\bar{X}_\alpha, \bar{X}_{\alpha'})}{[E(\tilde{\bar{X}}_s)]^2} + \frac{1 - P^m_s}{P^m_s n_s} \quad (16)$$

Since the first term in the right-hand side of the relation (16) is the relative variance of $\tilde{\bar{X}}_s$ under conditions that the $P'_{s\alpha}$ are not subject to sampling error, the effect of sampling error in the $P'_{s\alpha}$ on the relative variance of $\tilde{\bar{X}}_s$ would therefore be less than $(1 - P_s^m)/(P_s^m n_s)$, the second term in the right-hand side of relation (16). This is summarized in table C.

As is seen in table C, for all but one or two of the smallest States, the effect of sampling variance in the $P'_{s\alpha}$ on the relative variance of the synthetic estimator $\bar{X}_s$, would be quite small.

The approximate variance of $\tilde{\bar{X}}_{su}$, the synthetic estimator for subdomains, can be expressed in a form parallel to expression (14) as

$$\sigma^2_{\tilde{\bar{X}}_{su}} \doteq \sum_{\alpha \in u} \frac{P^2_{s\alpha}}{f^2_{su}} \sigma^2_{\bar{X}'_\alpha} + \frac{1}{n_s} \sum_{\alpha \in u} \bar{X}^2_\alpha \frac{P_{s\alpha}}{f_{su}} \left(1 - \frac{P_{s\alpha}}{f_{su}}\right) + 2 \sum_{\substack{\alpha < \alpha' \\ \alpha', \alpha \in u}} \frac{P_{s\alpha}}{f_{su}} \cdot \frac{P_{s\alpha'}}{f_{su}} \operatorname{Cov}(\bar{X}'_\alpha, \bar{X}'_{\alpha'}) \quad (17)$$

Sampling variances of synthetic estimates of HIS health characteristics for the 3-year period 1969-71 were obtained based on equations (14) and (17) with the following two modifications made to simplify the calculations:

1. $P_{s\alpha} \doteq \bar{P}_\alpha$ for all α, where P_α represents the proportion of the U.S. population in cell α. That is, the proportion of the population in any α cell is approximately the same for all States.

Table C. Maximum contribution to the relative variance of $\tilde{\bar{X}}_s$ of sampling variation in the $P_{s\alpha}$

n_s	P_s^m				
	.0001	.001	.01	.05	.10
	Maximum contribution to the relative variance of $\tilde{\bar{X}}_s$				
1,000	10.00	1.00	.10	.019	.009
10,000	1.00	.10	.01	.0019	.0009
100,000	0.10	.01	.001	.00019	.00009

2. $\operatorname{Cov}(\bar{X}'_\alpha, \bar{X}'_{\alpha'}) = 0$ for all $\alpha < \alpha'$, so that the third term of equations (14) and (17) drops out.

Under these assumptions equation (14) reduces to

$$\sigma^2_{\tilde{\bar{X}}_s} \doteq \sum_{\alpha=1}^{50} \bar{P}^2_\alpha \sigma^2_{\bar{X}'_\alpha} + \frac{1}{n_s} \sum_{\alpha=1}^{50} \bar{X}^2_\alpha P_\alpha (1 - \bar{P}_\alpha) \quad (18)$$

and equation (17) becomes

$$\sigma^2_{\tilde{\bar{X}}_{su}} \doteq \sum_{\alpha \in u} \left(\frac{\bar{P}_\alpha}{f_u}\right)^2 \sigma^2_{\bar{X}'_\alpha} + \frac{1}{n_s} \sum_{\alpha \in u} \bar{X}^2_\alpha \frac{\bar{P}_\alpha}{f_u} \left(1 - \frac{\bar{P}_\alpha}{f_u}\right) \quad (19)$$

where

$$f_u = \sum_{\alpha \in u} \bar{P}_\alpha .$$

Equations (18) and (19) were the expressions used to compute sampling errors for the estimates in the report.[13] Almost all the estimates had sampling errors that were very small relative to the size of the estimates themselves. The relative standard error (RSE), defined by

$$\text{RSE}(\tilde{\bar{X}}_s) = \sqrt{\frac{\sigma^2_{\tilde{\bar{X}}_s}}{(\tilde{\bar{X}}_s)^2}}$$

and

$$\text{RSE}(\tilde{\bar{X}}_{su}) = \sqrt{\frac{\sigma^2_{\tilde{\bar{X}}_{su}}}{(\tilde{\bar{X}}_{su})^2}},$$

was 5 percent or less for virtually all statistics in the report, even for the smallest States. The only important exceptions occurred for estimates of the proportion of persons in certain population subgroups who were unable to carry on major activity. The most variable subgroup was the

under-45 age group, where the RSE ranged from 7.4 percent for the entire United States to 10.4 percent for Alaska. The highest single RSE was 11.6 percent for white persons in Alaska. Although it may seem strange for State estimates to have such small sampling errors, these estimates were essentially weighted averages of national HIS estimates based on 3 years of data collection.

Bias of synthetic estimator.—The synthetic estimator is a biased estimator with the bias $B(\tilde{\bar{X}}_s)$ given by

$$B(\tilde{\bar{X}}_s) = \sum_{\alpha=1}^{k} P_{s\alpha} (\bar{X}_\alpha - \bar{X}_{s\alpha}) \qquad (20)$$

where $\bar{X}_{s\alpha}$ is the true mean level of characteristic X for demographic cell α in State s.

Regression-Adjusted Estimator

Background.—One of the basic limitations on the synthetic estimator $\tilde{\bar{X}}_s$ is that it is adjusted only for the specific set of demographic cells (or α cells) taken into consideration. If the parameter being estimated is influenced by variables other than those taken into consideration by the α cells, then the synthetic estimator will not reflect this influence. Often it is not possible to include in the α-cell array all the variables thought to be important because data on these variables are not available in sufficient demographic detail. However, although a particular variable might not be able to be used in the synthetic estimator, it can often be taken into consideration in other ways. In an earlier article,[7] a method was proposed to take into consideration such variables.

Method of estimation.—The method presented below uses the synthetic estimator $\tilde{\bar{X}}_s$ in conjunction with a set of ancillary variables $z_{s1}, \ldots, z_{sm}$ to produce an adjusted synthetic estimator. In particular, we assume the linear model given by

$$Y_s = \alpha + \beta_1 z_{s1} + \ldots + \beta_m z_{sm} + \leftarrow_s \qquad (21)$$

where Y_s, the percentage difference between the synthetic estimate $\tilde{\bar{X}}_s$ and the true value $\bar{X}_s$ of characteristic X for State s is given by

$$Y_s = \left(\frac{\bar{X}_s - \tilde{\bar{X}}_s}{\tilde{\bar{X}}_s} \right) 100,$$

$\leftarrow_s$ = term representing random error,

$z_{s1}, \ldots, z_{sm}$ = values of variables $z_1, \ldots, a_m$ for State s, and

$\alpha, \beta_1, \ldots, \beta_m$ = regression coefficients to be estimated

If estimates $\hat{\alpha}$ of α, $\hat{\beta}_1$ of β_1, ..., and $\hat{\beta}_m$ of β_m were available and substituted into the right-hand side of equation (21), algebraic manipulation would result in an estimator $\hat{\bar{X}}_s$ of $\bar{X}_s$ given by

$$\hat{\bar{X}}_s = \tilde{\bar{X}}_s \left[1 + 0.01 \left(\hat{\alpha} + \hat{\beta}_1 z_{s1} + \ldots + \hat{\beta}_m z_{sm} \right) \right] \qquad (22)$$

Equation (21) states that the percentage difference Y_s between the synthetic estimate $\tilde{\bar{X}}_s$ and the true value $\bar{X}_s$ is a linear function of a set of variables $z_{s1}, \ldots, z_{sm}$. For example, z_{s1} might be the proportion of persons in State s living in SMSA's; z_{s2}, the proportion of persons in State s having family income below the poverty level, and so forth. Equation (21) expresses the concept that except for random variation the percentage difference between a synthetic estimate and a true value is a linear function of a set of variables $z_{s1}, \ldots, z_{sm}$. The estimator given by equation (22) is called the regression-adjusted estimator, and it was used and evaluated by Levy[7] in computing State estimates of deaths from motor vehicles for the year 1960. In that study, it was found to be an improvement over the synthetic estimator. However, it can be used only when relevant ancillary data are available.

EVALUATION OF SYNTHETIC ESTIMATES

Background

A fundamental problem of the synthetic estimation procedure has been the difficulty in evaluating the estimates produced by this methodology. Although expressions have been derived for estimating the sampling variance of synthetic estimates, it is much more difficult to estimate the bias of a synthetic estimate, and since sampling errors are often small for synthetic estimates, the bias may often make the largest contribution to the total mean square error. A method has been developed, however, by investigators at the U.S. Bureau of the Census[9,14] for estimating the mean square error of synthetic estimates by somewhat indirect means.

Another consideration of importance in obtaining synthetic estimates is their sensitivity to the particular set of α cells used in producing them. Although a more detailed α-cell grid should produce synthetic estimates having lower bias, the potential reduction in bias may in fact be small and may not justify the cost of obtaining the detailed α-cell grid. This issue has been addressed in an empirical study using the synthetic estimates of disability, utilization of health services, and limitation of activity based on 1969-71 data from the Health Interview Survey and is discussed in a later section.

Estimation of Mean Square Error (MSE) and Average Mean Square Error (AMSE) of Synthetic Estimates

A procedure has been developed by Waksberg and Gonzalez[9] which enables the mean square error of a synthetic estimate to be estimated provided that an unbiased estimate of the same characteristic exists for the same population which is uncorrelated with the synthetic estimate. This procedure is developed by means of theorems presented below:

Theorem 6: Let $\tilde{\bar{X}}_s$ estimate a parameter $\bar{X}_s$ with bias given by $B(\tilde{\bar{X}}_s)$ and let $\bar{X}'_s$ be an unbiased estimate of $\bar{X}_s$ which is uncorrelated with $\tilde{\bar{X}}_s$. Then the following relation is true:

$$E(\bar{X}'_s - \tilde{\bar{X}}_s)^2 = \mathrm{MSE}_{\tilde{\bar{X}}_s} + \sigma^2_{\bar{X}'_s} \qquad (23)$$

where

$\mathrm{MSE}_{\tilde{\bar{X}}_s}$ = the mean square error of $\tilde{\bar{X}}_s$

and

$\sigma^2_{\bar{X}'_s}$ = the variance of $\bar{X}'_s$.

Theorem 7: If $\tilde{\bar{X}}_s$ is an estimate of $\bar{X}_s$ with bias given by $B(\tilde{\bar{X}}_s)$, if $\bar{X}'_s$ is an unbiased estimate of $\bar{X}_s$ uncorrelated with $\tilde{\bar{X}}_s$ and if $\hat{\sigma}^2_{\bar{X}'_s}$ is an unbiased estimate of $\sigma^2_{\bar{X}'_s}$, then the estimate $\hat{\mathrm{MSE}}_{\tilde{\bar{X}}_s}$ given by

$$\hat{\mathrm{MSE}}_{\tilde{\bar{X}}_s} = (\bar{X}'_s - \tilde{\bar{X}}_s)^2 - \sigma^2_{\bar{X}'_s} \qquad (24)$$

is an unbiased estimate of $\mathrm{MSE}_{\tilde{\bar{X}}_s}$.

Investigators at the Census Bureau have used the relationship given by equation (24) to evaluate synthetic estimates for certain variables such as unemployment where independent estimates are available. However, a serious limitation on the use of the estimated mean square error $\hat{\mathrm{MSE}}_{\tilde{\bar{X}}_s}$ as given in equation (24) is its likely instability since the unbiased estimate $\bar{X}'_s$ and the estimate $\hat{\sigma}^2_{\tilde{\bar{X}}_s}$ of its variance are both likely to have large variances themselves, since, in all likelihood, they would be based on relatively small sample size. Aware of this, Gonzalez and Waksberg have introduced the concept of evaluating synthetic estimates not by their individual mean square error, but by the average mean square error (AMSE) of a set of synthetic estimates. Specifically, the AMSE of a set of S synthetic estimates is given by

$$\mathrm{AMSE} = \frac{1}{S}\left[E \sum_{s=1}^{S} (\tilde{\bar{X}}_s - \bar{X}_s)^2 \right] \qquad (25)$$

and is estimated without bias by the expression, $\widehat{AMSE}$, given by

$$\widehat{AMSE} = \frac{1}{S}\sum_{s=1}^{S}(\hat{X}_s - X_s')^2 - \frac{1}{S}\sum_{s=1}^{S}\sigma^2_{X_s'}. \quad (26)$$

This statistic has been used with certain elaborations by Census Bureau investigators as the major criterion for evaluating synthetic estimates. A shortcoming of this criterion, however, is that it does not yield an estimate of the mean square error for a specific synthetic estimate (e.g., estimated unemployment in Ohio, 1976). Rather, it gives the *average* mean square error a set of synthetic estimates.

Evaluation of HIS Synthetic Estimates for Alternative α-Cell Grids

Investigators at NCHS originally hoped to evaluate the 1969-71 HIS synthetic estimates by means of the AMSE criterion. However, there was no unbiased estimate uncorrelated with the synthetic estimate that could be used in equation (26). Although it is thought that the bias of the nearly unbiased estimate discussed above is likely to be small for HIS variables, and that the correlation between the synthetic estimate and nearly unbiased estimate is also likely to be small, the task of obtaining a reasonable estimate of its variance is difficult since it is often based on data from one or two primary sampling units.

The main thrust in the evaluation of the 1969-71 HIS synthetic estimates was an empirical investigation comparing synthetic estimates based on the 50 α-cell grid used to obtain the published 1969-71 synthetic estimates with those obtained by collapsing the 50 cells into a smaller grid. In particular, synthetic estimates were obtained for the 50 States and the District of Columbia based on (1) the total 50 α-cell grid, (2) a 2 α-cell grid based only on sex, (3) a 4 α-cell grid based on age alone, (4) an 8 α-cell grid based on sex and age, (5) a 16 α-cell grid based on color, sex, and age, and (6) a 16 α-cell grid based on family income, sex, and age. The synthetic estimates produced by each of the collapsed grids (sex, age, sex by age, sex by age and color, and sex by age and income) were compared with the synthetic estimates produced by the total 50 α-cell grid by use of the following summary statistics:

1. The mean over all 50 States and the District of Columbia of the proportional absolute difference $\frac{|\hat{X}_{s,g} - \hat{X}_s|}{\hat{X}_s}$ between the synthetic estimate $\hat{X}_{s,g}$ based on a particular grid and the synthetic estimate $\hat{X}_s$ based on the total 50 α-cell grid (table D).
2. The maximum over all 50 States and the District of Columbia of the proportional absolute difference defined above (table E).
3. The correlation coefficient over all 50 States and the District of Columbia between the synthetic estimate produced by a collapsed grid with that produced by the total grid (table F).

The mean proportional absolute difference (table D) is a measure of the average relative difference between synthetic estimates produced by a collapsed grid and those produced by the total grid. For the HIS variables considered in this study, synthetic estimates produced by each of the collapsed grids agreed closely by this criterion with synthetic estimates produced by the detailed 50 α-cell grid. For most of the 14 varibles in this study, the mean proportional absolute difference was less than 5 percent, and the worst agreement by this criterion was shown by the synthetic estimates produced by the α-cell grid based on sex. In most cases, the synthetic estimates based on age by sex, age by sex and color, and age by sex and income did not show substantially better agreement by this criterion with those based on the total detailed grid than the synthetic estimates based on age alone.

The maximum proportional absolute difference (table E) gives a feeling of the extent that a single synthetic estimate based on a collapsed grid might differ from the comparable synthetic estimate based on the detailed 50-cell grid. The magnitudes of some of the statistics shown in table E imply that in individual States, the grid used to compute the synthetic estimates might affect the size of the estimate, even though the

Table D. Mean proportional absolute differences between the synthetic estimate produced by the total 50 α-cell grid and that produced by other α-cell grids for selected Health Interview Survey (HIS) variables, 1969-71

HIS variable	α-cell grid				
	Sex	Age	Sex by age	Sex by color by age	Sex by income by age
	Mean proportional absolute difference				
Restricted activity days	.028	.017	.017	.024	.020
Bed disability days	.029	.026	.025	.032	.021
Work loss days	.039	.026	.025	.032	.021
Hospital discharges per 100 person years	.021	.008	.008	.008	.008
Average length of hospitalization	.035	.034	.034	.028	.023
Percent of persons having one or more hospital episodes in a year	.018	.008	.008	.007	.006
Percent of persons having one or more physician visits in a year	.008	.008	.008	.005	.007
Number of physician visits per person year	.021	.016	.016	.015	.017
Percent of persons having one or more dental visits in a year	.031	.031	.031	.016	.027
Number of dental visits per person year	.056	.055	.055	.032	.062
Percent not limited in activity	.008	.003	.003	.003	.004
Percent limited in activity	.061	.021	.021	.029	.026
Percent limited in amount or kind of major activity	.064	.024	.024	.025	.027
Percent unable to carry on major activity	.064	.058	.061	.081	.057

Table E. Maximum proportional absolute differences between the synthetic estimate produced by the total 50 α-cell grid and that produced by other α-cell grids for selected Health Interview Survey (HIS) variables, 1969-71

HIS variable	α-cell grid				
	Sex	Age	Sex by age	Sex by color by age	Sex by income by age
	Maximum proportional absolute difference				
Restricted activity days	.103	.073	.073	.099	.136
Bed disability days	.134	.088	.089	.114	.170
Work loss days	.216	.047	.140	.071	.211
Hospital discharges per 100 person years	.127	.055	.070	.034	.024
Average length of hospitalization	.189	.201	.203	.083	.266
Percent of persons having one or more hospital episodes in a year	.103	.057	.073	.026	.020
Percent of persons having one or more physician visits in a year	.042	.041	.042	.019	.027
Number of physician visits per person year	.104	.084	.088	.060	.048
Percent of persons having one or more dental visits in a year	.221	.235	.237	.058	.241
Number of dental visits per person year	.382	393	.398	.101	.402
Percent not limited in activity	.037	.013	.013	.014	.021
Percent limited in activity	.454	.094	.092	.130	.184
Percent limited in amount or kind of major activity	.493	.088	.091	.122	.154
Percent unable to carry on major activity	.895	.233	.192	.264	.494

Table F. Correlation coefficients between the synthetic estimate produced by the total 50 α-cell grid and that produced by other α-cell grids for selected Health Interview Survey (HIS) variables, 1969-71

HIS variable	α-cell grid				
	Sex	Age	Sex by age	Sex by color by age	Sex by income by age
	Correlation coefficient				
Restricted activity days	.89	.96	.96	.94	.93
Bed disability days	.94	.96	.96	.95	.94
Work loss days	.81	.89	.90	.86	.85
Hospital discharges per 100 person years	.88	.98	.97	.99	.99
Average length of hospitalization	.89	.89	.89	.96	.87
Percent of persons having one or more hospital episodes in a year	.90	.98	.98	.99	.99
Percent of persons having one or more physician visits in a year	.81	.82	.81	.95	.91
Number of physician visits per person year	.93	.96	.96	.98	.98
Percent of persons having one or more dental visits in a year	.91	.90	.90	.98	.92
Number of dental visits per person year	.96	.95	.95	.99	.96
Percent not limited in activity	.83	.96	.96	.95	.92
Percent limited in activity	.61	.95	.95	.93	.93
Percent limited in amount or kind of major activity	.50	.94	.94	.94	.93
Percent unable to carry on major activity	.87	.94	.94	.93	.91

average proportional absolute differences are small.

The correlation coefficient between synthetic estimates based on the detailed 50-cell grid and those based on a less detailed grid (table F) measures the strength of the relationship between the two sets of estimates. It is a measure that is particularly appropriate when it is desired to rank a set of estimates and when the absolute values of the estimates are of secondary interest. In general, the correlations were quite high, with estimates based on the sex grid showing the lowest correlations with those based on the detailed grid.

———— ooo ————

REFERENCES

[1]National Center for Health Statistics: *Synthetic State Estimates of Disability.* PHS Publication No. 1759. Public Health Service. Washington. U.S. Government Printing Office, 1968.

[2]U.S. Bureau of the Census: *The Methods and Materials of Demography*, by H. S. Shryock, J. S. Siegel, and associates. Washington. U.S. Government Printing Office, 1973.

[3]U.S. Bureau of the Census: Developments in statistical estimation for local areas, by E. P. Ericksen. *Statistical Methodology of Revenue Sharing and Related Estimated Studies.* Census Tract Papers, Series GE-40, No. 10. Washington U.S. Government Printing Office, 1974. pp. 51-56.

[4]Ericksen, E. P.: A method for combining sample survey data and symptomatic indicators to obtain population estimates for local areas. *Demography* 10: 137-160, 1973.

[5]Ericksen, E. P.: A regression method for estimating population changes of local areas. *J. Am. Stat. Assoc.* 69: 867-875, 1974.

[6]Woodruff, R. S.: Use of a regression technique to produce area breakdowns of the monthly national estimates of retail trade. *J. Am. Stat. Assoc.* 61: 496-504, 1936.

[7]Levy, P. S.: The use of mortality data in evaluating synthetic estimates, in *Proceedings of the American Statistical Association 1971, Social Statistics Section.* Washington. American Statistical Association, 1974. pp. 328-331.

[8]Gonzalez, M. E., and Hoza, C.: Small Area Estimation of Unemployment. Presented at meeting of International Statistical Institute, Warsaw, Poland, 1975.

[9]Gonzalez, M. E., and Waksberg, J. E.: Estimation of the Error of Synthetic Estimates. Presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria, Aug. 1973.

[10]Simmons, W. R.: Adjustment of Data-Synthetic Estimates. Presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria, Aug. 1973.

[11]U.S. Bureau of the Census: Use and evaluation of synthetic estimates, by M. E. Gonzalez. *Statistical Methodology of Revenue Sharing and Related Estimated Studies.* Census Tract Papers, Series GE-40, No. 10. Washington. U.S. Government Printing Office. 1974. pp. 46-50.

[12]Namekata, T., Levy, P. S., and O'Rourke, T. W.: Synthetic estimates of work loss disability for each State and the District of Columbia. *Public Health Rep.* 90: 532-538, 1975.

[13]National Center for Health Statistics: *State Estimates of Disability and Utilization of Medical Services, United States, 1969-71.* DHEW Pub. No. (HRA) 77-1241. Health Resources Administration. Washington, U.S. Government Printing Office, Jan. 1977.

[14]Gonzalez, M. E.: Small area estimation of unemployment, in *Proceedings of the American Statistical Association, Social Statistics Section.* Washington. American Statistical Association, 1975.

[15]Hansen, M., Hurwitz, W., and Madow, W.: *Sample Survey Methods and Theory, Vol. 1.* Methods and applications. New York. John Wiley & Sons, Inc., 1953.

APPENDIX

Proofs of Lemmas and Theorems 18
Lemma 1 18
Lemma 2 18
Lemma 3 19
Theorem 4 19
Theorem 5 20
Theorem 6 21
Theorem 7 22

APPENDIX

PROOF OF LEMMAS AND THEOREMS

Lemma 1: The expectation $E(\bar{X}'_s)$ of the nearly unbiased estimator $\bar{X}'_s$ is given by

$$E(\bar{X}'_s) = \sum_{j=1}^{J} n_{sj.} \bar{X}_{.j.}/n_{s..} \qquad (1A)$$

where

$$\bar{X}_{.j.} = \sum_{s=1}^{S} n_{sj.} \bar{X}_{sj.}/n_{.j.}$$

= average level of X in stratum j,

$$n_{.j.} = \sum_{s=1}^{S} n_{sj.}$$

= the number of persons in stratum j,

$$\bar{X}_{sj.} = \sum_{i=1}^{I_{sj}} n_{sji} \bar{X}_{sji}/n_{sj.}$$

= average level of X in that portion of stratum j which is in State s,

and

$\bar{X}_{sji}$ = the average level of characteristic X in that portion of State s which is in PSU i of stratum j.

Proof

We note that the expectation of $\bar{X}'_s$ is give by

$$E(\bar{X}'_s) = E\left(\sum_{j=1}^{J} n_{sj.} \bar{X}'_j/n_{s..}\right)$$

$$= \sum_{j=1}^{J} n_{sj.} E(\bar{X}'_j)/n_{s..}$$

$$= \sum_{j=1}^{J} n_{sj.} \bar{X}_{.j.}/n_{s..}$$

since $\bar{X}'_j$ is an unbiased estimator of $\bar{X}_{.j.}$.

QED

Lemma 2: The bias $B(\bar{X}'_s)$ in the nearly unbiased estimate $\bar{X}'_s$ can be expressed by

$$B(\bar{X}'_s) = \sum_{j=1}^{J} n_{sj.} (\bar{X}_{.j.} - \bar{X}_{sj.})/n_{s..} \qquad (2A)$$

or equivalently by

$$B(\bar{X}'_s) = \sum_{j=1}^{J} \frac{n_{sj.}}{n_{s..}} \sum_{i=1}^{I_{sj}} \left(\frac{\bar{X}_{.j.}}{I_{sj}} - \frac{n_{sji}}{n_{sj.}} \bar{X}_{sji}\right). \qquad (3A)$$

Proof

Since, by definition, the bias $B(\bar{X}'_s)$ of $\bar{X}'_s$ is equal to

$$B(\bar{X}'_s) = E(\bar{X}'_s) - \bar{X}'_{s..}$$

where

$$\bar{X}_{s..} = \sum_{j=1}^{J} n_{sj.} \bar{X}_{sj.}/n_{s..}$$

= average level of X in State s, relation (2A) follows directly from lemma 1 and the definition of $\bar{X}_{s..}$.

We note that

$$\bar{X}_{.j.} - \bar{X}_{sj.} = \bar{X}_{.j.} - \sum_{i=1}^{I_{sj}} \frac{n_{sji}\bar{X}_{sji}}{n_{sj.}}$$

$$= \sum_{i=1}^{I_{sj}} \left(\frac{\bar{X}_{.j.}}{I_{sj}} - \frac{n_{sji}}{n_{sj.}} \bar{X}_{sji} \right) \qquad (4A)$$

Therefore, relation (3A) follows from relation (2A) by substitution of relation (4A) into relation (2A).

QED

Lemma 3: Let us assume that the ratio $n_{sji}/n_{sj.}$ is the same for all PSU's in the same stratum in the same State, which implies that

$$\frac{n_{sji}}{n_{sj.}} = \frac{1}{I_{sj}} \text{ for } i=1, \ldots, I_{sj}. \qquad (5A)$$

Then the bias $B(\bar{X}'_s)$ in the "nearly" unbiased estimate $\bar{X}'_s$ has the form given by

$$B(\bar{X}'_s) = \sum_{j=1}^{J} \frac{n_{sj.}}{n_{s..}} \frac{1}{I_{sj}} \sum_{i=1}^{I_{sj}} (\bar{X}_{.j.} - \bar{X}_{sji}). \qquad (6A)$$

Proof

The results follow directly from relation (5A) and lemma 2.

QED

Theorem 4: If $\frac{n_{sji}}{n_{sj.}} = \frac{1}{I_{sj}}$ for $i=1, \ldots, I_{sj}$, then $B^2(\bar{X}'_s)$, the square of the bias in $\bar{X}'_s$, is given by the expression

$$B^2(\bar{X}'_s) = \sum_{j=1}^{J} \frac{n_{sj.}^2}{n_{s..}^2} \frac{\sigma_{sji}^2}{I_{sj}} + \sum_{j=1}^{J} \frac{n_{sj.}^2}{n_{s..}^2} \frac{1}{I_{sj}} (\bar{X}_{sj.} - \bar{X}_{.j.})^2 + 2 \sum_{j=1}^{J} \frac{n_{sj.}^2}{n_{s..}^2} \frac{1}{I_{sj}^2} \sum_{i=1}^{I_{sj}} \sum_{i'<i} (\bar{X}_{.j.} - \bar{X}_{sji}) (\bar{X}_{.j.} - \bar{X}_{sji'})$$

$$+ 2 \sum_{j=1}^{J} \sum_{k<j} \frac{n_{sj.} n_{sk.}}{n_{s..}^2 I_{sj} I_{sk}} \sum_{i=1}^{I_{sj}} (\bar{X}_{.j.} - \bar{X}_{sji}) \sum_{i=1}^{I_{sk}} (\bar{X}_{.k.} - \bar{X}_{ski}) \qquad (7A)$$

---

where

$$\sigma_{sji}^2 = \frac{1}{I_{sj}} \sum_{i=1}^{I_{sj}} (\bar{X}_{sji} - \bar{X}_{sj.})^2.$$

Proof

By adding and subtracting $\bar{X}_{sj.}$ to the right-hand side of relation (6A) we obtain

$$B(\bar{X}'_s) = \sum_{j=1}^{J} \frac{n_{sj.}}{n_{s..}} \frac{1}{I_{sj}} \times$$

$$\sum_{i=1}^{I_{sj}} (\bar{X}_{.j.} - \bar{X}_{sj.} + \bar{X}_{sj.} - \bar{X}_{sji}). \qquad (8A)$$

Squaring the right-hand side of relation (8A), we obtain

$$B^2(\bar{X}'_s) = \left\{\sum_{j=1}^{J} \frac{n_{sj.}}{n_{s..}} \frac{1}{I_{sj}} \sum_{i=1}^{I_{sj}} [(\bar{X}_{.j.} - \bar{X}_{sj.}) + (\bar{X}_{sj.} - \bar{X}_{sji})]\right\}^2$$

$$= \sum_{j=1}^{J} \frac{n_{sj.}^2}{n_{s..}^2} \frac{1}{I_{sj}^2} \left\{\sum_{i=1}^{I_{sj}} [(\bar{X}_{.j.} - \bar{X}_{sj.}) + (\bar{X}_{sj.} - \bar{X}_{sji})]\right\}^2 + 2 \sum_{j=1}^{J} \sum_{k<j} \frac{n_{sj.} n_{sk.}}{n_{s..}^2 I_{sj} I_{sk}} \left[\sum_{i=1}^{I_{sj}} (\bar{X}_{.j.} - \bar{X}_{sji})\right]\left[\sum_{i=1}^{I_{sk}} (\bar{X}_{.k.} - \bar{X}_{skl})\right]$$

$$= \sum_{j=1}^{J} \frac{n_{sj.}^2}{n_{s..}^2} \frac{1}{I_{sj}^2} \sum_{i=1}^{I_{sj}} (\bar{X}_{.j.} - \bar{X}_{sj.} + \bar{X}_{sj.} - \bar{X}_{sji})^2 + 2 \sum_{j=1}^{J} \frac{n_{sj.}^2}{n_{s..}^2} \frac{1}{I_{sj}^2} \sum_{i=1}^{I_{sj}} \sum_{i'<i} (\bar{X}_{.j.} - \bar{X}_{sji})(\bar{X}_{.j.} - \bar{X}_{sji'})$$

$$+ 2 \sum_{j=1}^{J} \sum_{k<j} \frac{n_{sj.} n_{sk.}}{n_{s..}^2 I_{sj} I_{sk}} \sum_{i=1}^{I_{sj}} (\bar{X}_{.j.} - \bar{X}_{sji}) \sum_{i=1}^{I_{sk}} (\bar{X}_{.k.} - \bar{X}_{skl}). \qquad (9A)$$

But the first term in equation 9A

$$\sum_{j=1}^{J} \frac{n_{sj.}^2}{n_{s..}^2} \frac{1}{I_{sj}^2} \sum_{i=1}^{I_{sj}} (\bar{X}_{.j.} - \bar{X}_{sj.} + \bar{X}_{sj.} - \bar{X}_{sji})^2$$

$$= \sum_{j=1}^{J} \frac{n_{sj.}^2}{n_{s..}^2} \frac{1}{I_{sj}^2} \left[\sum_{i=1}^{I_{sj}} (\bar{X}_{sj.} - \bar{X}_{.j.})^2 + \sum_{i=1}^{I_{sj}} (\bar{X}_{sji} - \bar{X}_{sj.})^2\right]$$

$$= \sum_{j=1}^{J} \frac{n_{sj.}^2}{n_{s..}^2} \frac{1}{I_{sj}} [(\bar{X}_{sj.} - \bar{X}_{.j.})^2 + \sigma_{sji}^2]. \qquad (10A)$$

Therefore, by appropriate substitution of relations (10A) into (9A), we obtain the form specified by equation (7A).

QED

Theorem 5: If the $P'_{s\alpha}$ are independent of the $\bar{X}'_\alpha$, then the variance of $\bar{\bar{X}}_s$ given by

$$\sigma^2_{\bar{\bar{X}}_s} = \mathrm{Var}\left(\sum_{\alpha=1}^{k} P'_{s\alpha} \bar{X}'_\alpha\right)$$

$$= \sum_{\alpha=1}^{k} \mathrm{Var}\,(P'_{s\alpha} \bar{X}'_\alpha)$$

$$+ 2 \sum_{\alpha<\alpha'} \mathrm{Cov}\,(P'_{s\alpha} \bar{X}'_\alpha, P'_{s\alpha'} \bar{X}'_{\alpha'}) \qquad (11A)$$

reduces to

$$\sigma^2_{\bar{\bar{X}}_s} \doteq \sum_{\alpha=1}^{k} P^2_{s\alpha} \sigma^2_{\bar{X}'_\alpha} + \frac{1}{n_s} \sum_{\alpha=1}^{k} \bar{X}^2_\alpha (1 - P_{s\alpha}) P_{s\alpha}$$

$$+ 2 \sum_{\alpha<\alpha'} P_{s\alpha} P_{s\alpha'} \mathrm{Cov}\,(\bar{X}'_\alpha, \bar{X}'_{\alpha'}). \qquad (12A$$

for large values of n

Proof

The second term in the right-hand side of equation (11A) is given by

$$\text{Cov}\,(P'_{s\alpha}\,\bar{X}'_{\alpha}, P'_{s\alpha'}\,\bar{X}'_{\alpha'})$$

$$= E(P'_{s\alpha}\,\bar{X}'_{\alpha}\,P'_{s\alpha'}\,\bar{X}'_{\alpha'}) - E(P'_{s\alpha}\,\bar{X}'_{\alpha})E(P'_{s\alpha'}\,\bar{X}'_{\alpha'})$$

$$= E(P'_{s\alpha}\,P'_{s\alpha'})\,E(\bar{X}'_{\alpha}\,\bar{X}'_{\alpha'})$$

$$- E(P'_{s\alpha})E(P'_{s\alpha'})E(\bar{X}'_{\alpha})E(\bar{X}'_{\alpha'}) \quad (13A)$$

since $P'_{s\alpha}$ is independent of $\bar{X}'_{\alpha}$.

But

$E(P'_{s\alpha}) = P_{s\alpha}$ (the true proportion of State s falling into cell α), and

$$\text{Cov}\,(P'_{s\alpha}, P'_{s\alpha'}) = \frac{-P_{s\alpha}\,P_{s\alpha'}}{n_s}. \quad (14A)$$

where

n_s = the sample size in State s used for estimating the $P_{s\alpha}$ (e.g., for a State having 1 million persons and for the 1-Percent Public Use Sample Tapes, $n_s = 10{,}000$).

Therefore,

$$E(P'_{s\alpha}\,P'_{s\alpha'}) = \text{Cov}\,(P'_{s\alpha}, P'_{s\alpha'}) + E(P'_{s\alpha})\,E(P'_{s\alpha'})$$

$$\doteq \frac{-P_{s\alpha}\,P_{s\alpha'}}{n_s} + P_{s\alpha}\,P_{s\alpha'}$$

$$= \frac{n_s - 1}{n_s} P_{s\alpha}\,P_{s\alpha'} \quad (15A)$$

and for large n_s

$$E(P'_{s\alpha}\,P'_{s\alpha'}) \doteq P_{s\alpha}\,P_{s\alpha'}. \quad (16A)$$

Therefore, from equations (13A) and (16A)

$$\text{Cov}\,(P'_{s\alpha}\,\bar{X}'_{\alpha}, P'_{s\alpha'}\,\bar{X}'_{\alpha'}) = P_{s\alpha}\,P_{s\alpha'}\,[E(\bar{X}'_{\alpha}\,\bar{X}'_{\alpha'})$$

$$- E(\bar{X}'_{\alpha})E(\bar{X}'_{\alpha'})] = P_{s\alpha}\,P_{s\alpha'}\,\text{Cov}(\bar{X}'_{\alpha}, \bar{X}'_{\alpha'}). \quad (17A)$$

Hansen, Hurwitz, and Madow[15] show that

$$\text{Var}\,(P'_{s\alpha}\,\bar{X}'_{\alpha}) \doteq P^2_{s\alpha}\,\bar{X}^2_{\alpha}$$

$$\times\left[\frac{\text{Var}\,P'_{s\alpha}}{P^2_{s\alpha}} + \frac{\text{Var}\,\bar{X}'_{\alpha}}{\bar{X}^2_{\alpha}} + \frac{2\,\text{Cov}\,(\bar{X}'_{\alpha}, P'_{s\alpha})}{\bar{X}_{\alpha}\,P_{s\alpha}}\right]$$

which reduces to

$$\text{Var}(P'_{s\alpha}\,\bar{X}'_{\alpha}) \doteq P^2_{s\alpha}\,X^2_{\alpha}\left[\frac{\text{Var}\,P'_{s\alpha}}{P^2_{s\alpha}} + \frac{\text{Var}\,\bar{X}'_{\alpha}}{\bar{X}^2_{\alpha}}\right] \quad (18A)$$

Since

$$\text{Cov}\,(\bar{X}'_{\alpha}, P'_{s\alpha}) = 0$$

and this reduces further to a form given by

$$\text{Var}\,(P'_{s\alpha}\,\bar{X}'_{\alpha}) \doteq \bar{X}^2_{\alpha}\,\text{Var}\,P'_{s\alpha} + P^2_{s\alpha}\,\text{Var}\,\bar{X}'_{\alpha}$$

$$= \frac{\bar{X}^2_{\alpha}\,P_{s\alpha}\,(1 - P_{s\alpha})}{n_s} + P^2_{s\alpha}\,\sigma^2_{\bar{X}'_{\alpha}}.$$

Substituting equations (17A) and (19A) into equation (11A),

$$\sigma^2_{\bar{\bar{X}}'_s} \doteq \sum_{\alpha=1}^{k} P^2_{s\alpha}\,\sigma^2_{\bar{X}'_{\alpha}} + \frac{1}{n_s}\sum_{\alpha=1}^{k} \bar{X}^2_{\alpha}\,(1 - P_{s\alpha})\,P_{s\alpha}$$

$$+ 2\sum_{\alpha<\alpha'} P_{s\alpha}\,P_{s\alpha'}\,\text{Cov}\,(\bar{X}'_{\alpha}, \bar{X}'_{\alpha'}).$$

QED

Theorem 6: Let $\tilde{\bar{X}}_s$ estimate a parameter $\bar{X}_s$ with bias given by $B(\tilde{\bar{X}}_s)$ and let $\bar{X}'_s$ be an unbiased estimate of $\bar{X}_s$ which is uncorrelated with $\tilde{\bar{X}}_s$. Then the following relation is true

$$E(\bar{X}'_s - \tilde{\bar{X}}_s)^2 = \text{MSE}_{\tilde{\bar{X}}_s} + \sigma^2_{\bar{X}'_s} \quad (20A)$$

NOTE: A list of references follows the text.

where

$MSE_{\tilde{\bar{X}}_s}$ = the mean square error of $\tilde{\bar{X}}_s$

and

$\sigma^2_{\bar{X}'_s}$ = the variance of $\bar{X}'_s$.

Proof

$$E(\bar{X}'_s - \tilde{\bar{X}}_s)^2 = E\,[(\bar{X}'_s - \bar{X}_s) + (\bar{X}_s - \tilde{\bar{X}}_s)]^2$$

$$= E(\bar{X}'_s - \bar{X}_s)^2 + E(\bar{X}_s - \tilde{\bar{X}}_s)^2$$

$$- 2\,E(\bar{X}'_s - \bar{X}_s)\,(\tilde{\bar{X}}_s - \bar{X}_s)$$

but

$$E(\bar{X}'_s - \bar{X}_s)^2 = \sigma^2_{\bar{X}'_s}$$

and

$$E(\bar{X}_s - \tilde{\bar{X}}_s)^2 = MSE_{\tilde{\bar{X}}_s}.$$

Also, it can be shown that

$$E(\bar{X}'_s - \bar{X}_s)\,(\tilde{\bar{X}}_s - \bar{X}_s) = \mathrm{Cov}\,(\bar{X}'_s, \tilde{\bar{X}}_s)$$

$$+ B(\bar{X}'_s)\,B(\tilde{\bar{X}}_s) = 0$$

since $\bar{X}'_s$ and $\tilde{\bar{X}}_s$ are uncorrelated and $B(\bar{X}'_s) = 0$.

QED

Theorem 7: If $\tilde{\bar{X}}_s$ is an estimate of $\bar{X}_s$ with bias given by $B(\tilde{\bar{X}}_s)$, if $\bar{X}'_s$ is an unbiased estimate of $\bar{X}_s$ uncorrelated with $\tilde{\bar{X}}_s$, and if $\hat{\sigma}^2_{\bar{X}'_s}$ is an unbiased estimate of $\sigma^2_{\bar{X}'_s}$, then the estimate $\widehat{MSE}_{\tilde{\bar{X}}_s}$ given by

$$\widehat{MSE}_{\tilde{\bar{X}}_s} = (\bar{X}'_s - \bar{X}_s)^2 - \hat{\sigma}^2_{\bar{X}'_s} \qquad (21A)$$

is an unbiased estimate of $MSE_{\tilde{\bar{X}}_s}$.

Proof

Proof follows directly from theorem 6.

QED

——— ○○○ ———

★U.S. GOVERNMENT PRINTING OFFICE: 1977—260-837:36

VITAL AND HEALTH STATISTICS PUBLICATIONS SERIES

Formerly Public Health Service Publication No. 1000

Series 1. Programs and Collection Procedures.—Reports which describe the general programs of the National Center for Health Statistics and its offices and divisions, data collection methods used, definitions, and other material necessary for understanding the data.

Series 2. Data Evaluation and Methods Research.—Studies of new statistical methodology including experimental tests of new survey methods, studies of vital statistics collection methods, new analytical techniques, objective evaluations of reliability of collected data, contributions to statistical theory.

Series 3. Analytical Studies.—Reports presenting analytical or interpretive studies based on vital and health statistics, carrying the analysis further than the expository types of reports in the other series.

Series 4. Documents and Committee Reports.—Final reports of major committees concerned with vital and health statistics, and documents such as recommended model vital registration laws and revised birth and death certificates.

Series 10. Data from the Health Interview Survey.—Statistics on illness; accidental injuries; disability; use of hospital, medical, dental, and other services; and other health-related topics, based on data collected in a continuing national household interview survey.

Series 11. Data from the Health Examination Survey.—Data from direct examination, testing, and measurement of national samples of the civilian, noninstitutionalized population provide the basis for two types of reports: (1) estimates of the medically defined prevalence of specific diseases in the United States and the distributions of the population with respect to physical, physiological, and psychological characteristics; and (2) analysis of relationships among the various measurements without reference to an explicit finite universe of persons.

Series 12. Data from the Institutionalized Population Surveys.—Discontinued effective 1975. Future reports from these surveys will be in Series 13.

Series 13. Data on Health Resources Utilization.—Statistics on the utilization of health manpower and facilities providing long-term care, ambulatory care, hospital care, and family planning services.

Series 14. Data on Health Resources: Manpower and Facilities.—Statistics on the numbers, geographic distribution, and characteristics of health resources including physicians, dentists, nurses, other health occupations, hospitals, nursing homes, and outpatient facilities.

Series 20. Data on Mortality.—Various statistics on mortality other than as included in regular annual or monthly reports. Special analyses by cause of death, age, and other demographic variables; geographic and time series analyses; and statistics on characteristics of deaths not available from the vital records, based on sample surveys of those records.

Series 21. Data on Natality, Marriage, and Divorce.—Various statistics on natality, marriage, and divorce other than as included in regular annual or monthly reports. Special analyses by demographic variables; geographic and time series analyses; studies of fertility; and statistics on characteristics of births not available from the vital records, based on sample surveys of those records.

Series 22. Data from the National Mortality and Natality Surveys.—Discontinued effective 1975. Future reports from these sample surveys based on vital records will be included in Series 20 and 21, respectively.

Series 23. Data from the National Survey of Family Growth.—Statistics on fertility, family formation and dissolution, family planning, and related maternal and infant health topics derived from a biennial survey of a nationwide probability sample of ever-married women 15-44 years of age.

For a list of titles of reports published in these series, write to:

Scientific and Technical Information Branch
National Center for Health Statistics
Public Health Service
Hyattsville, Md. 20782